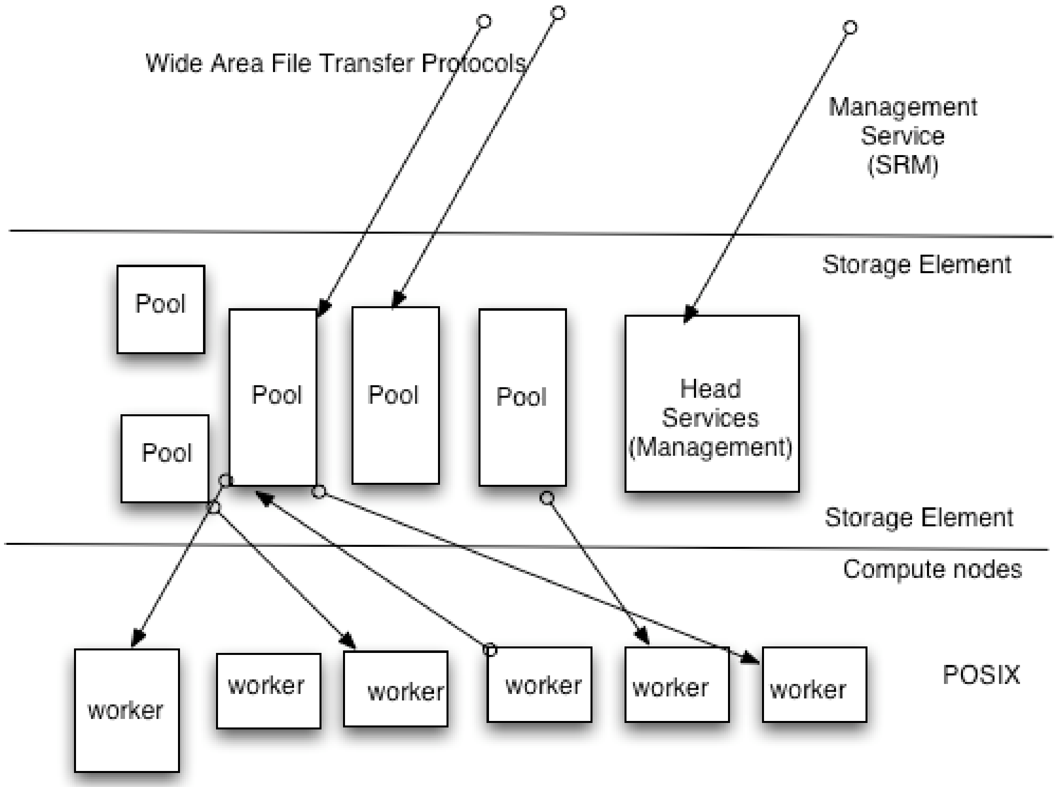
Storage, Networking and Data Management

Don Petravick, Fermilab

OSG Milwaukee meeting

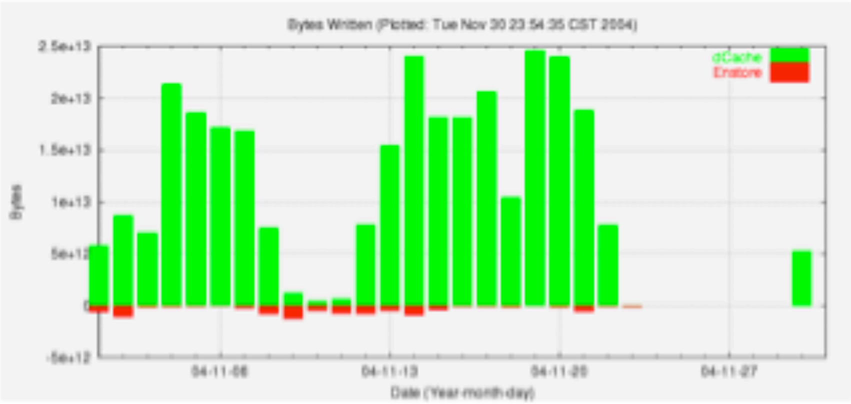
July, 2005

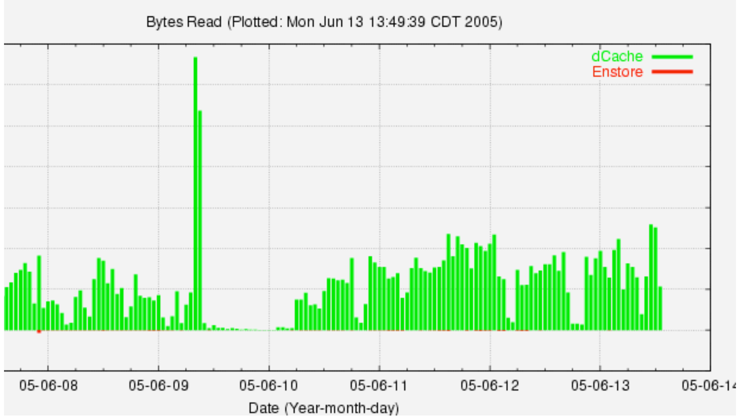
Storage Element



Open Science Grid

SE and SC's

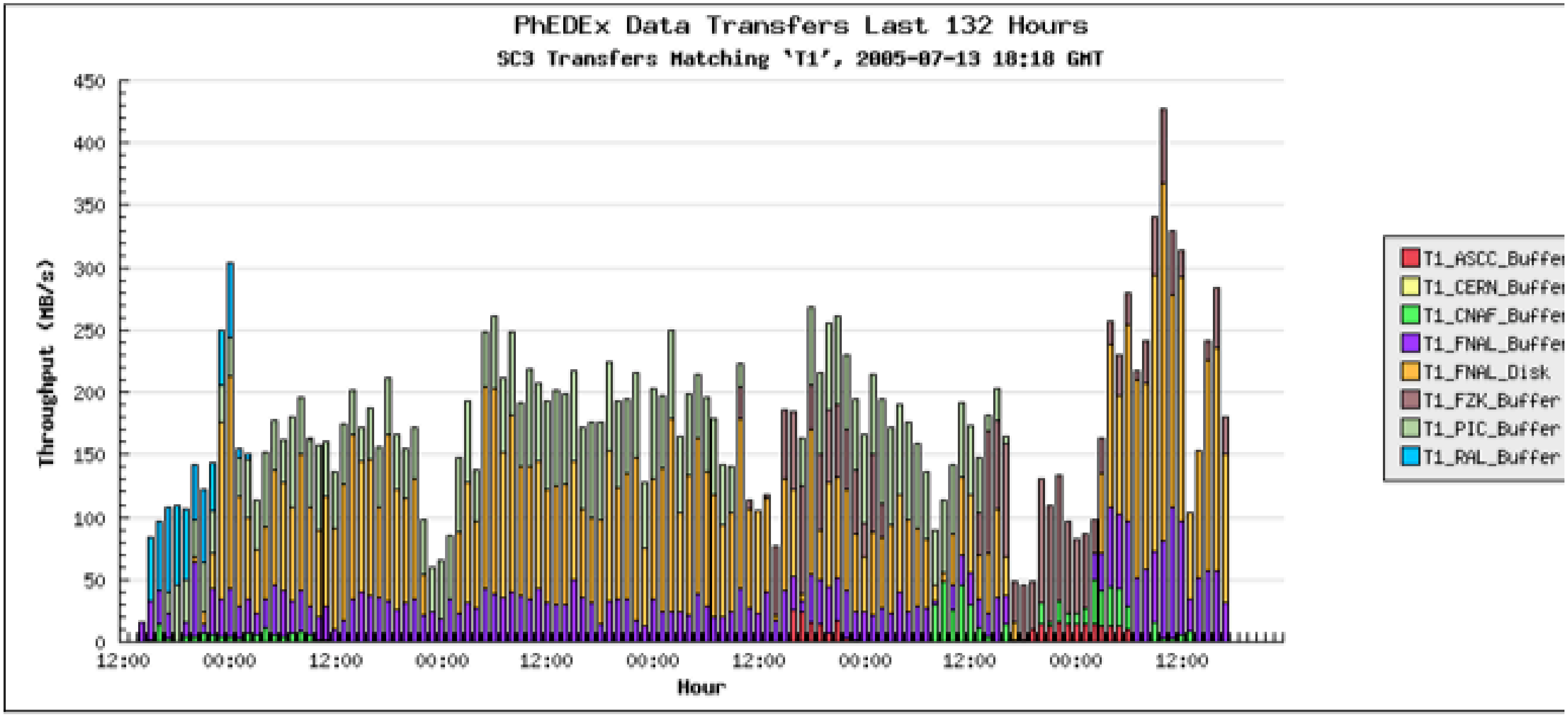




- Performance on local and grid Sides.
- Ease of use shock (if clusters are large error amplifiers, what are storage clusters)
- Can a large number of groups do this?

Achievements

- Drilling down into the better parts (T1)



Site Service Choices

Tier 0/1s

- CERN
  - Storage: Castor-2/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL LFC Oracle
  - Does CERN participate as T1?
- FNAL
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL Globus RLS
- CNAF
  - Storage: Castor-1/SRM
  - Transfer: PhEDEx/SRM (srmcp)
  - File catalogue: POOL LFC Oracle
- RAL
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL LFC Oracle
- CCIN2P3
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL LFC Oracle
- PIC
  - Storage: Castor-1/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL LFC MySQL
- FZK
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL LFC Oracle
- ASCC
  - Storage: Castor(-1?)/SRM
  - Transfers: PhEDEx/SRM (srmcp)?
  - File catalogue: POOL LFC Oracle

Site Service Choices

Tier 2s

- US: Florida, Wisconsin, San Diego, Caltech (+ Purdue, Nebraska, MIT?)
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL (POOL Globus RLS later at some?)
- Italy: Legnaro
  - Storage: LCG DPM/SRM
  - Transfer: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL
- Spain: CIEMAT
  - Storage: Castor-1/SRM
  - Transfer: PhEDEx/SRM (srmcp) (Globus as fallback)
  - File catalogue: POOL MySQL
- UK: Imperial
  - Storage: dCache/SRM
  - Transfer: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL
- Germany: DESY
  - Storage: dCache/SRM (+ tape)
  - Transfer: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL (?)
- France: ?
- Taiwan: ?

What We Really Used

Tier 0/1s

- CERN
  - Storage: Castor-1/SRM
  - Transfers: None
  - File catalogue: textfile + grep
- FNAL
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL
- CNAF
  - Storage: Castor-1/SRM / SE
  - Transfer: PhEDEx/SRM / globus-url-copy
  - File catalogue: POOL MySQL
- RAL
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL
- CCIN2P3
  - (No CMS transfers so far)
- PIC
  - Storage: Castor-1 (not SRM)
  - Transfers: PhEDEx/Globus (globus-url-copy)
  - File catalogue: POOL MySQL
- FZK
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL
- ASCC
  - Storage: Castor-1/SRM
  - Transfers: PhEDEx/Globus (globus-url-copy)
  - File catalogue: POOL MySQL
- All
  - Production networks

What We Really Used

Tier 2s

- US: Purdue, Nebraska, Wisconsin (Florida, San Diego, Caltech)
  - Storage: dCache/SRM
  - Transfers: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL (POOL Globus RLS later at some?)
  - No transfers at FL, UCSD, Caltech
- Italy: Legnaro
  - Storage: LCG DPM/SRM
  - Transfer: PhEDEx/Globus
  - File catalogue: POOL MySQL
  - Transfers for one day
- Spain: CIEMAT / IFCA
  - Storage: Castor-1/SRM
  - Transfer: PhEDEx/SRM (srmcp) (Globus as fallback)
  - File catalogue: POOL MySQL
- UK: Imperial
  - Storage: dCache/SRM
  - Transfer: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL
- Germany: DESY
  - Storage: dCache/SRM (+ tape)
  - Transfer: PhEDEx/SRM (srmcp)
  - File catalogue: POOL MySQL
- France: ?
- Taiwan: ?

OSG Deployment Status

- Gridcat shows one SE on the OSG.
  - USCMS T1 facility at FNAL.
- Not on GridCAT, but existing
  - BNL Atlas T1 has SE software
  - Six CMS T2's have SEs (3 in current SC)
  - Perhaps not a comprehensive list.
- FNAL General facility soon.

SE software status

- Dcache (FNAL/DESY)
  - Accepted by US CMS, US Atlas T1 facility, outside the US.
  - Support not available to general OSG sites.
  - Acceptance at LCG t1 and T2 sites.
- DRM (LBNL)
  - Evaluated at Florida in OSG context.
  - In the VDT.
  - Used by STAR/Evaluated at Florida
- CASTOR
- DPM -- From Europe. -- Used in the US?

Tour of SE issues

- Still -- Scalable IO with minimal data movement. (Deployment and use errors)
- Management of space.
- Fault mitigation
- Micro-scheduling
  - Management of spindle.
  - Management of socket resources
  - Management of local network congestion.
- Application role in firewall traversal
  - Firewalls and NAT's

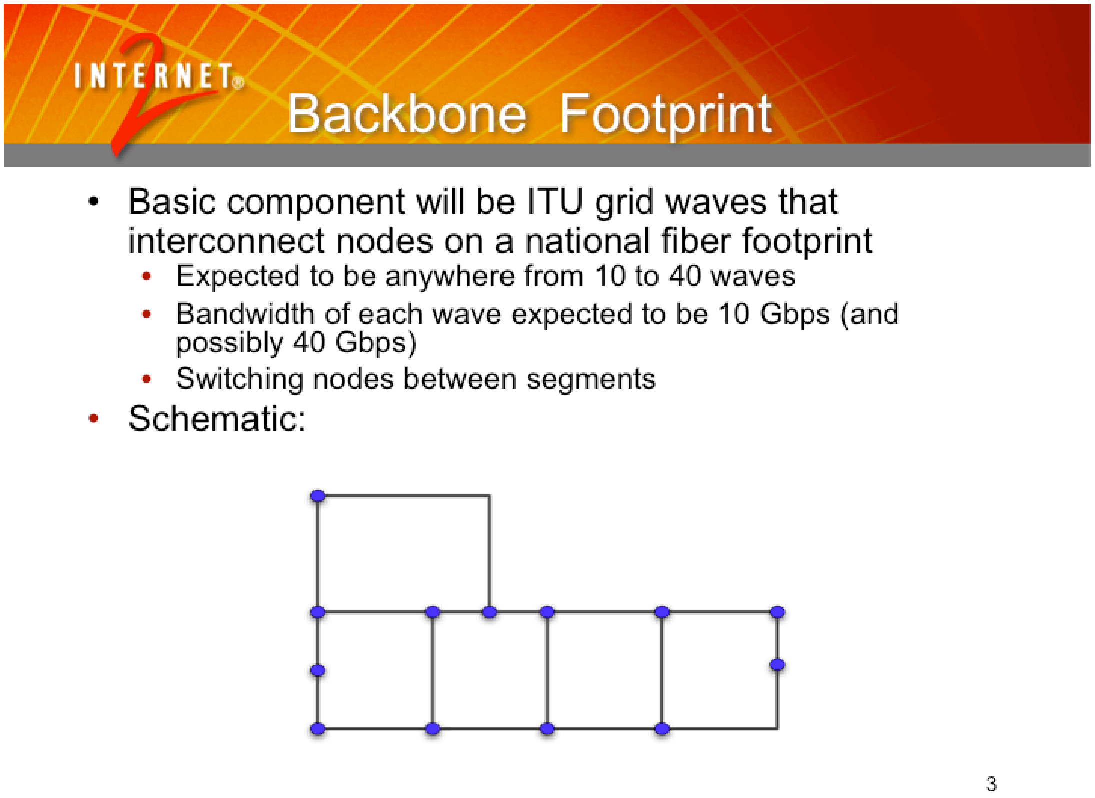
Software -- or -- Is The Perfect the enemy of the Good.

- Many claims of replicated services.
  - What help is our integration framework?
  - RFT <-> SRMCP
  - Name space
- Standardized Stanzas of web services v.s. standardized web services

US R&E Networking

- ESNet, Abilene loom large
- Available “open” infrastructure
  - Fiber Infrastructure Projects.
  - Open Model Optical Exchanges.
  - Initiatives integrating the above.
- Network and systems research involving custom provisioning.
- Practical initiatives by various institutions.
- Various last mile problems.

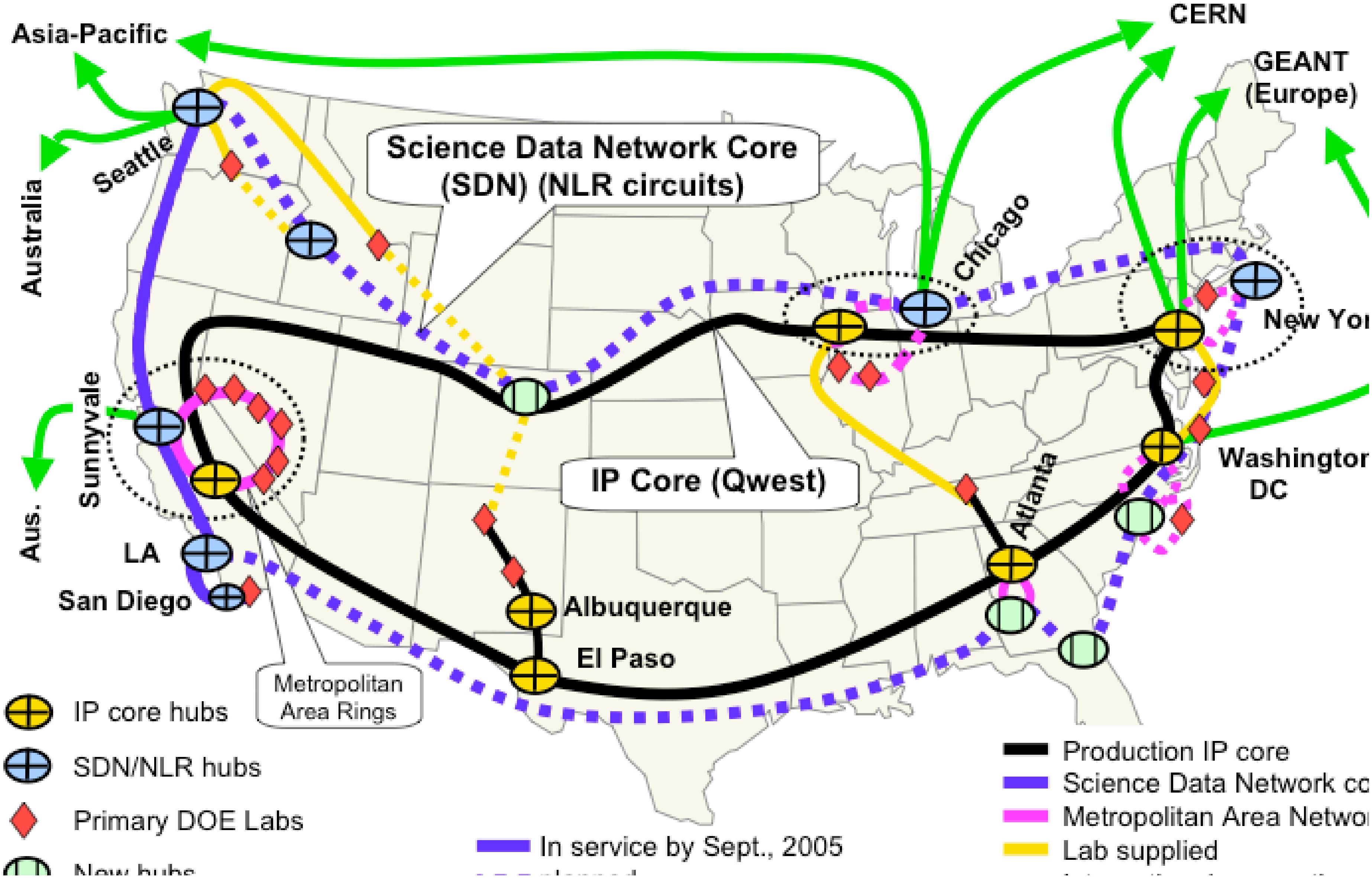
Internet2



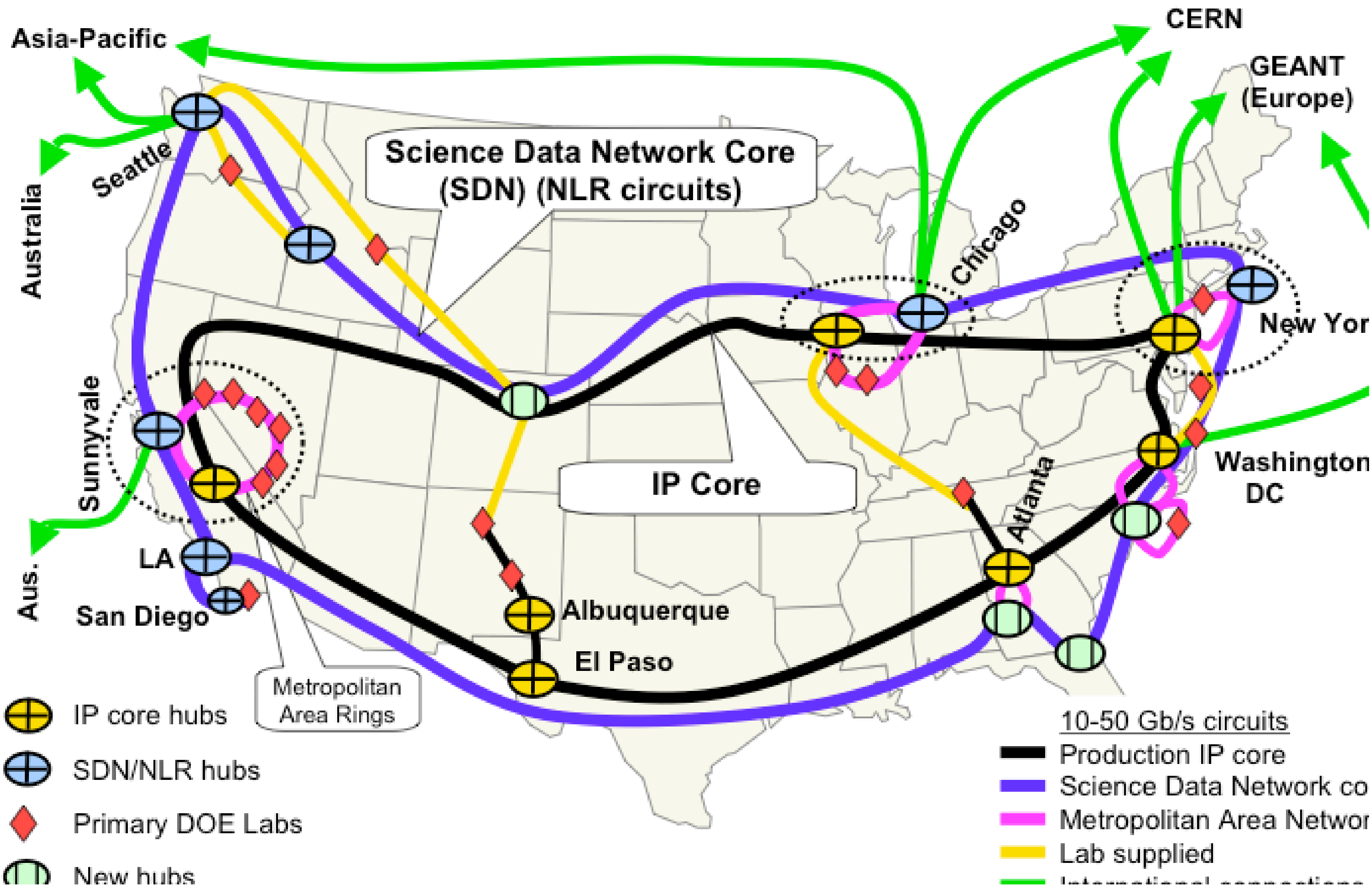
INTERNET2 Backbone Footprint

- Basic component will be ITU grid waves that interconnect nodes on a national fiber footprint
  - Expected to be anywhere from 10 to 40 waves
  - Bandwidth of each wave expected to be 10 Gbps (and possibly 40 Gbps)
  - Switching nodes between segments
- Schematic:

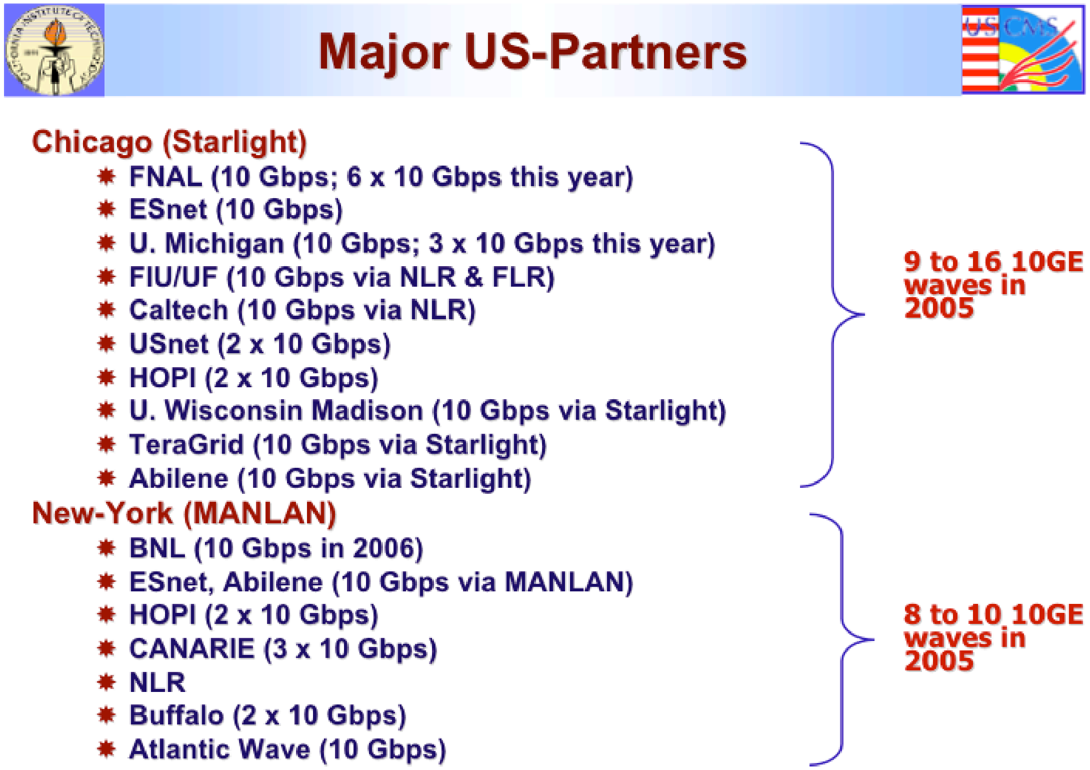
Evolution of ESnet – Step One:
SF Bay Area MAN and West Coast SDN



ESnet Target Architecture:
IP Core+Science Data Network Core+Metro Area Rings



Other Waves (LHCnet partners)



Major US-Partners

Chicago (Starlight)

- FNAL (10 Gbps; 6 x 10 Gbps this year)
- ESnet (10 Gbps)
- U. Michigan (10 Gbps; 3 x 10 Gbps this year)
- FIU/UF (10 Gbps via NLR & FLR)
- Caltech (10 Gbps via NLR)
- USnet (2 x 10 Gbps)
- HOPI (2 x 10 Gbps)
- U. Wisconsin Madison (10 Gbps via Starlight)
- TeraGrid (10 Gbps via Starlight)
- Abilene (10 Gbps via Starlight)

9 to 16 10GE waves in 2005

New-York (MANLAN)

- BNL (10 Gbps in 2006)
- ESnet, Abilene (10 Gbps via MANLAN)
- HOPI (2 x 10 Gbps)
- CANARIE (3 x 10 Gbps)
- NLR
- Buffalo (2 x 10 Gbps)
- Atlantic Wave (10 Gbps)

8 to 10 10GE waves in 2005

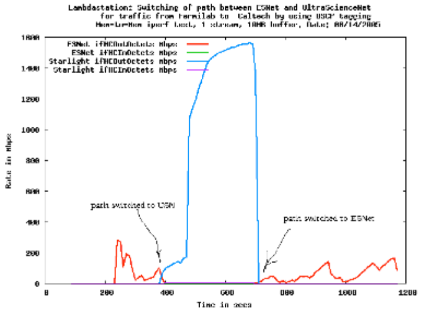
Alternate Paths, Production

- Important to link data centers to the wide area properly.
- OSG can play a role in achieving an overall integration that is useful for our VO’s

LambdaStation Project: Dynamic flows based switching between ESNet and UltraScienceNet.

Prepared by Andrey Bobyshev, bobyshev@fnal.gov 06/15/2005

The graph below represents the results of throughput measurement for selective switching of flows between two alternative paths. The tests were conducted between Fermilab and Caltech for memory to memory data transferring by using ***iperf*** tool, 1 stream, 10MB buffer size. Path switching is based on policy based routing configured at both sites for pairs of source/destination addresses marked by DSCP. Path switching is initiated on host site by turning DSCP tagging on or off for specific flows. In our tests we used ***iptables*** to do actual tagging for traffic assosiated with certain UID (user identifier) or PID (process identifier). The test was started via ESNet path with five streams. Throughput has reached about 350Mbps. To avoid saturation of the OC12 link at Fermilab the number of streams was reduced to only one. In about 3 mins hosts on both ends started DSCP tagging of traffic to switch forwarding via USN path, and then, in about 5 mins tagging was switched off to forward packets via ESNet path again.



Summary(1)

OSG is to deliver a grid from components that are mature enough to integrate.

Storage

- Deliver on notion of storage element.
- Debate (at the blueprint level about modified storage element)
- Maturing software and delivery in the context of LHC
- Integrated performance can be considered quite good, but not near potential.

Summary(2)

- A networking monitoring program has begun.
- There examples of the creation and use of advanced infrastructures
  - DISUN/Ultralight model is there,and supported by NSF. Relied on by the HEP program.
- Its is also important to do as well as we can to exploit well-provisioned conventional networks.
  - Grid framework should be apropos to Network provisioning plans.

Summary (3)

- Data Management
  - Need to foster belief that Wide area connections and associated storage elements are
    - More available
    - More performant
    - And are implemented well on the OSG and its partners.
- This is possible, but is separate work from technology development.
  - Hardening of technology
  - Demonstration by production people
  - Management of expectations.